НЕКОТОРЫЕ ОСОБЕННОСТИ, ВОЗНИКАЮЩИЕ ПРИ ИЗУЧЕНИИ НЕЛИНЕЙНОЙ РЕГРЕССИИ С ИСПОЛЬЗОВАНИЕМ EXCEL И ДРУГИХ ПРОГРАММ

УДК 330.43
ВАК 08.00.13
РИНЦ 1818-4243

Ирина Владленовна Орлова,
к.э.н., профессор, профессор каф. Моделирование экономических и информационных систем Финансового университета при Правительстве РФ
Тел. (499) 277-21-44
Эл. почта: IVOrlova@fa.ru

Виктор Борисович Турундаевский,
к.э.н., доцент, профессор, каф. Прикладной математики Московского государственного университета экономики, статистики и информатики (МЭСИ)
Тел. (495) 442-60-98
Эл. почта: vik_turund@mail.ru

В статье излагаются методические и алгоритмические особенности, возникающие при моделировании экономических процессов с помощью моделей нелинейной регрессии. Обсуждается вопросы, связанные с разными подходами к оценке параметров нелинейной регрессии. Даются рекомендации для построения более качественных моделей нелинейной регрессии.

Ключевые слова: нелинейная регрессия, метод наименьших квадратов, степенная модель, линеаризация.

Irina Vladlenovna Orlova,
PhD in Economics, Professor, Professor, the Department of Modeling of economic and informational systems, the Financial University under the Government of the Russian Federation
Tel. (499) 277-21-44
E-mail: IVOrlova@fa.ru

Viktor Borisovich Turundaevskiy,
PhD in Economics, Associate Professor, Professor, the Department of Applied mathematics, the Moscow state University of Economics, statistics and Informatics (MESI)
Tel. (495) 442-60-98
E-mail: vik_turund@mail.ru

SOME PECULIARITIES ARISING IN THE STUDY OF NONLINEAR REGRESSION USING EXCEL AND OTHER PROGRAMS

The article considers the methodological and algorithmic features arising in modeling economic processes with the help of nonlinear regression models. Questions related to different approaches to estimation of nonlinear regression parameters are discussed in the article. Recommendations to build better models of non-linear regression are made.

Keywords: nonlinear regression, least-squares method, exponential model, linearization.

1. При рассмотрении зависимости экономических показателей на основе реальных статистических данных с использованием аппарата теории вероятности и математической статистики можно сделать выводы, что линейные зависимости не всегда адекватно описывают исследуемые процессы. Линейные зависимости рассматриваются лишь как частный случай для удобства и наглядности рассмотрения изучаемого экономического процесса. Чаще встречаются модели, которые отражают экономические процессы в виде нелинейной зависимости.

Например: затухающие гармонические и не гармонические колебания, которые могут характеризовать объемы продаж сезонного товара на этапе ухода с рынка. Или зависимости, характеризующие с экономической точки зрения жизненный цикл товара, ремаркетинг или конверсию товара.

Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций.

Изучение нелинейных моделей регрессии является обязательной частью программы дисциплины эконометрика. Как известно, преподавание дисциплины «Эконометрика» ведется исходя из требований, установленных в федеральном государственном образовательном стандарте высшего профессионального образования (ФГОС ВПО) и обязательных при реализации основных образовательных программ бакалавриата и магистратуры по направлению подготовки 080100 «Экономика» с использованием современного программного обеспечения. Рассмотрим некоторые проблемы, возникающие при изучении нелинейной регрессии с использованием Excel и других программных продуктов, и наметим пути их решения.

2. Рассмотрим статистическую связь между зависимой переменной y (случайная функция) и независимой переменной x (фактор) в следующем виде

$$V(y) = A + B \cdot U(x) + \delta, \tag{1}$$

где $V(y)$, $U(x)$ – известные монотонные функции,
A и B – неизвестные параметры (постоянные);
δ – случайная функция, $M\delta = 0$, $D(\delta) = \sigma^2$.

В точках: $x_1, x_2, ..., x_n$ известны значения зависимой переменной $y_1, y_2, ..., y_n$, а следовательно, можно вычислить

$$U_1 = U(x_1),\ U_2 = U(x_2),\ ...,\ U_n = U(x_n);$$
$$V_1 = V(y_1),\ V_2 = V(y_2),\ ...,\ V_n = V(y_n).$$

и записать

$$V_i = A + BU_i + \delta_i,\ i = 1, 2, ..., n.$$

Предположим, что $M(\delta_i\delta_j) = 0$, т.е. δ_i, δ_j – некоррелированные и одинаково распределенные случайные величины. Тогда методом наименьших квадратов (МНК) можно получить оценки $\hat{A}$ и $\hat{B}$ точных значений коэффициентов A и B уравнения (1):

$$\hat{B} = \frac{\overline{VU} - \overline{V} \cdot \overline{U}}{\overline{U^2} - \overline{U}^2},\quad \hat{A} = \overline{U} - \hat{B}\overline{U},$$

где $\overline{VU}, \overline{V}, \overline{U}, \overline{U^2}$ – средние значения соответствующих функций, например,

$$\overline{U} = \frac{1}{n}\sum_{i=1}^{n} U(x_i),\quad \overline{VU} = \frac{1}{n}\sum_{i=1}^{n} V(y_i)U(x_i).$$

Функция (линейная регрессия)

$$\hat{V}(y) = \hat{A} + \hat{B}U(x)$$

будет оценкой среднего значения (математического ожидания) случайной функции $V(y)$.

При сделанных предположениях относительно δ можно найти несмещенную оценку S_δ^2 дисперсии $\sigma^2 = D(\delta_i)$:

$$S_\delta^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(V\left(y_i\right) - \hat{V}\left(y_i\right)\right)^2.$$

Оценки $\hat{A}, \hat{B}$, а также $\hat{V}\left(y_i\right)$ являются случайными величинами, причем

$$M\left(\hat{A}\right) = A,\ M\left(\hat{B}\right) = B,$$

$$M\left(\hat{V}\left(y_i\right)\right) = M\left(V\left(y_i\right)\right).$$

Оценки S_A^2, S_B^2, S_V^2, их дисперсий $D(\hat{A})$, $D(\hat{B})$, $D(\hat{V})$ находятся по формулам

$$S_B^2 = \frac{S_\delta^2}{n}\cdot\frac{1}{\overline{U^2} - \overline{U}^2}, S_A^2 = \overline{U^2}S_B^2,$$

$$S_{V(y)}^2 = \frac{S_\delta^2}{n}\left(1 + \frac{\left(U\left(x\right) - \overline{U}\right)^2}{\overline{U^2} - \overline{U}^2}\right).$$

Оценки средних квадратических отклонений S_B, S_A, $S_{V(y)}$ характеризуют ошибки определения истинных значений соответствующих величин A, B, $V(y)$.

Предположим дополнительно, что случайные величины $\delta_1, \delta_2, ..., \delta_n$ имеют нормальное распределение, тогда можно построить интервальные оценки A, B, $V(y)$, а именно, с вероятностью $1 - \alpha$ (α – уровень значимости) будут выполняться неравенства

$$\begin{gathered}\hat{A} - t_\alpha S_A \le A \le \hat{A} + t_\alpha S_A,\\ \hat{B} - t_\alpha S_B \le B \le \hat{B} + t_\alpha S_B,\\ \hat{V}(y) - t_\alpha S_\delta \Delta \le V(y) \le \hat{V}(y) + t_\alpha S_\delta \Delta,\end{gathered} \qquad (2)$$

где

$$\Delta = \sqrt{1 + \frac{1}{2}\left(1 + \frac{\left(U\left(x\right) - \overline{U}\right)^2}{\overline{U^2} - \overline{U}^2}\right)},$$

и t_α – критическое значение t – распределения Стьюдента для уровня значимости α и числа степеней свободы $n - 2$.

Для проверки значимости параметров A и B вычисляются статистики

$$t_A = \hat{A} / S_A,\ t_B = \hat{B} / S_B.$$

Если $\left|t_A\right| < t_\alpha$, то коэффициент Δ не значим, аналогичное правило действует для параметра B.

Регрессионную модель в виде (1) обычно получают при линеаризации нелинейных моделей [1], [2], [3]. Если исходная модель нелинейная по независимой переменной x

$$y = a + bf(x) + \varepsilon,$$

то принимаем $V(y) = y$, $U(x) = f(x)$, $\delta = \varepsilon$, находим непосредственно коэффициенты $\hat{a} = \hat{A}$, $\hat{b} = \hat{B}$; они являются несмещенными, наилучшими линейными оценками истинных значений параметров a и $в$.

Для некоторых моделей, нелинейных по параметрам, удается с помощью соответствующих преобразований привести их к виду (1). Например, для мультипликативных степенной и показательной моделей

$$y = ax^b \cdot \varepsilon,\ y = ab^x \cdot \varepsilon$$

путем логарифмирования получаем модели

$$\ln y = \ln a + b\ln x + \ln\varepsilon,$$

$$\ln y = \ln a + \ln b \cdot x + \ln\varepsilon,$$

совпадающие с (1). Для степенной модели будут найдены коэффициенты $\hat{A} = \ln\hat{a}$, $\hat{B} = \hat{b}$. Имеем: $M\hat{b} = M\hat{B} = b$, т.е. $\hat{b}$ – несмещенная оценка b. Далее $M\hat{A} = A = \ln a$, $\hat{A}$ – несмещенная оценка $\ln a$, но $\hat{a} = \exp(\hat{A})$ не будет несмещенной оценкой a и найденные оценки $\hat{a}$ и $\hat{b}$ не дают возможности построить точное уравнение регрессии y и x.

Если рассматривается, например, аддитивная степенная модель

$$y = ax^b + \varepsilon,\ y_i = ax_i^b + \varepsilon_i,\ i = 1, 2, ..., n,$$

то необходимо применять прямой метод вычисления параметров, т.е. находить минимум функции $Q(a, b)$:

$$Q\left(a, b\right) = \sum_{i=1}^{n}\left(y_i - ax_i^b\right)^2 \qquad (3)$$

численным методом. Оценки $\hat{a}$, $\hat{b}$, найденные по линеаризированному уравнению, т.е. как минимум функции $Q_1(a, b)$:

$$Q_1\left(a, b\right) = \sum_{i=1}^{n}\left(\ln y_i - A - b\ln x_i\right)^2,$$

могут заметно отличаться от оценок, найденных по первому способу. Покажем это на примере.

Рассматривалась модель

$$y_i = 1{,}5x_i^{1,2} + \varepsilon_i,\ i = 1, 2, ..., 16.$$

Величины $\varepsilon_1, \varepsilon_2, ..., \varepsilon_n$ моделировались как элементы случайной выборки из нормального распределения: $M\varepsilon_i = 0$, $D\varepsilon_i = 4$. Найденные значения y_i представлены в таб. 1.

Построение степенной модели с помощью Мастера диаграмм в Excel дает следующие результаты (рис. 1).

Используя для построения степенной модели прологарифмированные данные, методом наименьших квадратов (Анализ данных в Excel) получим: $\ln a = -0{,}3005$ $b = 1{,}6904$, после операции потенцирования получаются те же параметры: $\hat{a} = 0{,}5006$, $\hat{b} = 1{,}6904$. Однако анализируя ряд остатков (табл. 2), приходим к выводу, что **коэффициент детерминации в Мастере диаграмм для степенной модели выдается неверно**.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2} = \frac{508{,}866}{2311{,}854} = 0{,}7799.$$

Таблица 1

x_i	1	2	3	4	5	6	7	8
y_i	0,1	5,09	4,64	9,25	9,93	14,58	17,14	19,7
x_i	9	10	11	12	13	14	15	16
y_i	20,29	20,2	25,34	26,36	33,87	32,7	38,78	40,68

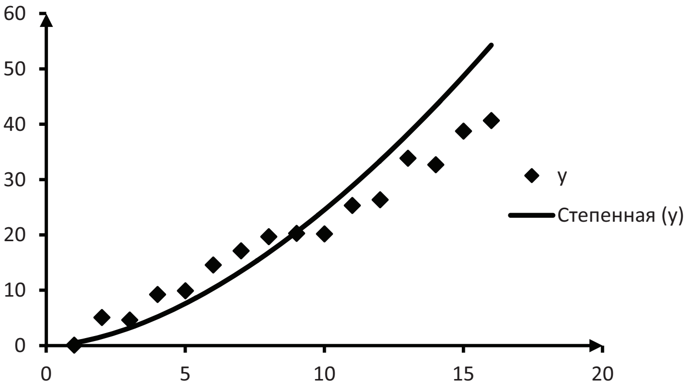

Рис. 1. График степенной модели $y = 0{,}5006x^{1{,}6904}$, $R^2 = 0{,}8425$

Таблица 2

Предсказанное	0,501	1,616	3,206	5,214	7,603	10,347	13,428	16,828
Остатки	–0,401	3,474	1,434	4,036	2,327	4,233	3,712	2,872

20,535	24,538	28,827	33,395	38,233	43,336	48,696	54,309
–0,245	–4,338	–3,487	–7,035	–4,363	–10,636	–9,916	–13,629

Таблица 3

Предсказанное	1,667	3,691	5,876	8,172	10,554	13,007	15,522	18,090
Остатки	–1,567	1,399	–1,236	1,078	–0,624	1,573	1,618	1,610
	20,705	23,363	26,061	28,795	31,562	34,361	37,189	40,045
	–0,415	–3,163	–0,721	–2,435	2,308	–1,661	1,591	0,635

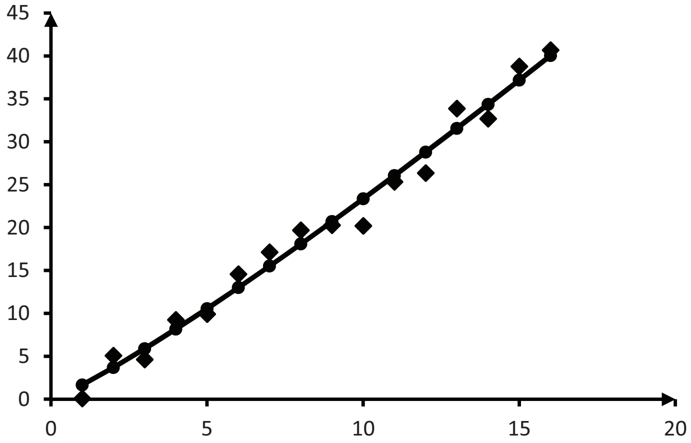

Рис. 1. График степенной модели $y = 1{,}673x^{1{,}145}$, $R^2 = 0{,}916$

Прямым методом наименьших квадратов, реализуя формулу (3) в Excel в Поиске решения было найдено: $\hat{a} = 1{,}673$, $\hat{b} = 1{,}145$. В табл. 3 приведены значения, предсказанные по модели $y = 1{,}673x^{1{,}145}$ и остатки. Коэффициент детерминации, вычисленный, на основании остатков (табл. 3), равен $R^2 = 0{,}916$. Как видим, качество модели хорошее, $Me = 0$, критерий Дарбина-Уотсона равен 2,3. График модели, приведенный на рис. 2 подтверждает наши выводы относительно этой модели.

Сделаем вывод, что при построении моделей нелинейной регрессии с помощью программных продуктов, полезно знать, каким способом выполняется оценка параметров модели.

Например, в SPSS [5] применяется прямой метод вычисления параметров, т.е. определяется минимум функции $Q(a,b) = \sum_{i=1}^{n}\left(y_i - ax_i^b\right)^2$. В программе VSTAT [6] оценки $\hat{a}$, $\hat{b}$, находятся по линеаризированному уравнению, т.е. как минимум функции

$$Q_1(a,b) = \sum_{i=1}^{n}\left(\ln y_i - A - b\ln x_i\right)^2.$$

3. При исследовании реальных нелинейных моделей возникают как аддитивные, так и мультипликативные модели.

Рассмотрим нелинейную мультипликативную модель в виде

$$y = \varepsilon\varphi(a, b, x),$$
$$(y_i = \varepsilon_i\varphi(a, b, x_i),\ M\varepsilon_i = 1),$$

где $\varphi(a, b, x)$ – известная функция своих аргументов. Для построения критерия метода наименьших квадратов запишем эти уравнения следующим образом:

$$y_i = \varphi(a, b, x_i) + (\varepsilon_i - 1)\varphi(a, b, x_i),$$
$$My_i = \varphi(a, b, x_i),$$

а затем

$$\frac{y_i}{\varphi(a,b,x_i)} - 1 = \varepsilon_i - 1,$$

отсюда получаем

$$Q(a,b) = \sum_{i=1}^{n}\left(\frac{y_i - \varphi(a,b,x_i)}{\varphi(a,b,x_i)}\right)^2 =$$

$$= \sum_{i=1}^{n}\left(\varepsilon_i - 1\right)^2. \qquad (4)$$

Для определения оценок a и b надо найти минимум этого критерия, например, численным методом. Так, для степенной модели будем иметь

$$Q(a,b) = \sum_{i=1}^{n}\left(\frac{y_i - ax_i^b}{ax_i^b}\right)^2. \qquad (5)$$

Возьмем степенную модель в виде

$$y_i = 1{,}5x_i^{1{,}2} + \varepsilon_i,\ i = 1, 2, ..., 16,$$

где ε_i – случайная величина, имеющая равномерное распределение на отрезке [0,8;1,2].

Значения случайной величины ε_i получены программой, описанной в [4], вычисленные значения y_i для заданных x_i представлены в таб. 4.

Путем минимизации критерия (5) было найдено: $\hat{a} = 1{,}622$, $\hat{b} = 1{,}157$, по линеаризированной модели получилось: $\hat{a} = 1{,}554$, $\hat{b} = 1{,}168$; оба результата практически совпадают и близки к истинным значениям этих коэффициентов. **Таким образом, для оценки параметров нелинейных мультипликативных моделей, не поддающихся линеаризации, можно использовать метод оптимизации по критерию (4).** Заметим, что при оптимизации по критерию (3) получились более грубые оценки: $\hat{a} = 2{,}154$, $\hat{b} = 1{,}024$. Однако именно этот подход реализован в программе SPSS.

При линеаризации модели возникает проблема восстановления исходной модели. Возьмем, например, обратную модель:

$$y = \frac{1}{a + bx + \varepsilon};$$

записывая

$$V(y) = \frac{1}{y} = a + bx + \varepsilon$$

получаем линейную модель, оценки $\hat{a}$, $\hat{b}$ параметров которой, полученные методом наименьших квадратов, являются несмещенными оценками параметров a, b исходной модели. Но уравнение

$$\hat{y} = \frac{1}{\hat{a} + \hat{b}x}$$

не является уравнением регрессии для случайной функции y, т.к.

$$\frac{1}{a + bx} \neq M\left(\frac{1}{a + bx + \varepsilon}\right).$$

Пусть для линеаризированной модели построен доверительный интервал (2). Считая $V(y)$ монотонно возрастающей функцией y, можем записать эквивалентное неравенство

$$V^{-1}(V(y) - t_\alpha S_\delta\Delta) \le y \le V^{-1}(V(y) + t_\alpha S_\delta\Delta) \quad (6)$$

т.е. доверительный интервал с уровнем значимости α для величины y. В качестве среднего значения $\hat{y}$ можно взять полусумму границ доверительного интервала:

Таблица 4

x_i	1	2	3	4	5	6	7	8
y_i	1,267	3,668	5,232	9,316	12,360	12,858	17,250	19,905
x_i	9	10	11	12	13	14	15	16
y_i	17,974	23,619	23,070	27,824	34,578	28,634	31,753	39,018

$$\hat{y} = \frac{1}{2}\left[\begin{array}{l} V^{-1}\left(\hat{V}\left(y\right) - t_\alpha S_\delta \Delta\right) \le y \le \\ \le V^{-1}\left(\hat{V}\left(y\right) + t_\alpha S_\delta \Delta\right) \end{array}\right].$$

Если $V(y)$ – монотонно убывающая функция, то в (6) знаки неравенств следует поменять на обратные. Для обратной модели, например, будем иметь:

$$\hat{y} = \frac{1}{\left(\hat{a} + \hat{b}x\right)\left[1 - \left(\frac{t_\alpha S_\alpha \Delta}{\hat{a} + \hat{b}x}\right)^2\right]}.$$

4.

В заключении следует отметить, что теоретический материал дисциплины «Эконометрика», читаемый в вузах, уменьшается, но его надо не только знать, но и обязательно понимать, чтобы грамотно строить и оценивать модели. Предполагается, что изучивший этот курс сможет решать практические задачи с использованием компьютера (тем более, что спектр эконометрических программных продуктов велик), ему не обязательно знать детально используемый в программе математический аппарат. Задачи решаются на компьютерах, но грамотное построение модели, понимание выдаваемых компьютером результатов и принятие на их основе решений остаётся за человеком, принимающим решение.

Литература

1. Эконометрика: учебник / под ред. В.С. Мхитаряна. – М.: Проспект, 2008. – 384 с.

2. Эконометрика: Учебник для магистров / Под. ред. И.И. Елисеевой. – М.: Издательство Юрайт, 2012. 453 с.

3. Орлова И.В., Половников В.А. Экономико-математические методы и модели: компьютерное моделирование: Учебное пособие – 3-е изд., перераб. и доп. ааа

4. Козлов А.Ю., Шишов В.Ф. Применение пакета анализа MS Excel в экономико-статистических расчетах. – М.: ЮНИТИ-ДАНА, 2003. – 139 с.

5. Бююль А., Цефель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей: Пер. с нем./Ахим Бююль, Петер Цефель. – СПб.:ООО «ДиаСофтЮП»,2002. – 608 с.

6. http://www.v-stat.ru/(дата обращения 18.09.2013).

7. Дрейпер, Норман, Смит, Гарри. Прикладной регрессионный анализ, 3-у изд. : Пер с англ. – М.: Издательский дом «Вильямс», 2007. – 912 с.

References:

1. Econometrics: textbook. /edited by V.S. Mhitaryan. – M.: Prospekt, 2008. – 384 s.

2. Econometrics: textbook for masters. / edited by Eliseeva I.I. – M.: Izdatelstvo Yurajt, 2012. 453 s.

3. Orlova I.V., Polovnikov V.A. Economic and mathematical methods and models: computer modeling. Textbook. M.: Vuzovskij uchebnik: INFRA-M, 2012. – 389 s.

4. Kozlov A.Yu., Shishov V.F. The use of Excel analysis package for economic-statistical estimation. M.: YuNITI-DANA, 2003.-139 s.

5. Byuyul A., Cefel P. SPSS: the art of information processing. Analysis of the statistic data and recovery of the latent consistencies. Per. s nem./ Ahim Byuyul, Peter Cefel. – SPb.:OOO «DiaSoftYuP»,2002. – 608 s.

6. http://www.v-stat.ru/(дата обращения 18.09.2013).

7. Drejper, Norman, Smith, Harry. Applied regression analysis. Per s angl. – M.: Izdatelskij dom «Vilyams», 2007. – 912 s.